Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp

Catherine A. Charneski & Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, Somerset, United Kingdom, BA2 7AY

Corresponding author: c.a.charneski@bath.ac.uk

Abstract

In the great majority of genomes the use of positive charge increases, on average, approaching protein N-termini. Such charged residues interacting with the negatively-charged exit tunnel slow ribosomes. This has been proposed to be selectively advantageous as it provides an elongation speed "ramp" at translational starts. Positive charges, however, are known to orientate proteins in membranes via the "positive inside rule" whereby excess charge lies on the cytoplasmic side of the membrane. Which of these two models better explains the N-terminal loading of positively charged amino acids? We find strong evidence that the tendency for average positive charge use to increase at termini is due exclusively to membrane protein topology: 1) increasing N-terminal positive charge is not found in cytosolic proteins, but in transmembrane ones with cytosolic N-termini, with signal sequences contributing additional charge; 2) positive charge density at N-termini corresponds to the length of cytoplasmically exposed transmembrane tails, its usage increasing just up until the membrane; 3) membrane-related patterns are repeated at C-termini, where no ramp is expected; 4) N-terminal positive charge patterns are no different from those seen internally in proteins in membrane-associated domains. The overall apparent increase in positive charge across all N-termini results from membrane proteins using positive charge adjacent to the cytosolic leaflet, combined with a skewed distribution of where N-termini cross the plasma membrane; 5) while *E. coli* was predicted to have a 5' ribosomal occupancy ramp of at least 31 codons, in contrast to what is seen in yeast, we find in ribosomal footprinting data no evidence for such a ramp. In sum, we find no need to invoke a translational ramp to explain the rising positive charge densities at N-termini. The membrane orientation model makes a full account of the trend.

Introduction

Why are some amino acids, or classes of amino acid, differentially distributed within proteins? Consider, for example, the location of positively charged residues. Enrichment of positive charge nearing protein N-termini has been demonstrated in humans (Berezovsky et al. 1999), *Escherichia coli* (Berezovsky et al. 1999) and *Saccharomyces cerevisiae* (Berezovsky et al. 1999; Tuller et al. 2011). Note that while the increase in use of positive charge nearing N-termini is true on average in a given genome, it does not follow that all proteins in any given genome have positive charges in this area. A successful model should then be able to explain why some proteins do and some do not contribute to the pattern of increasing positive charge use nearing the N-terminus. Here we consider two models that might explain this enrichment of positively charged amino acids at the starts of proteins.

The first model conjectures the positive charge enrichment is part of a "ramp" that controls ribosomal flow (Tuller et al. 2011). Positively charged amino acids are thought to be one (Lu et al. 2007; Tuller et al. 2011), possibly the principal (Charneski and Hurst 2013), determinant of ribosome velocity. The interior of the ribosomal exit tunnel is negatively charged (Lu et al. 2007) and positively charged residues within a protein are conjectured to interact with the negative charge in this channel, slowing ribosomal movement along transcripts (Lu and Deutsch 2008). This can explain, for example, why insertion of a long run of positively charged residues into a coding sequence stalls ribosomes (Ito-Harashima et al. 2007; Dimitrova et al. 2009). An excess of ribosomal density at the extreme 5' ends of transcripts is present in at least one dataset which profiled the location of ribosomes along transcripts (Ingolia et al. 2009; Tuller et al. 2010). As the extent of this enrichment correlates with, amongst other features, the density of charged amino acids, it has been proposed that the increase in charge at the N-termini of proteins exists as one part of a speed "ramp" to control the flow of ribosomes at the start of translation, possibly to somehow prevent downstream traffic jams between them (Tuller et al. 2011).

While the translational ramp may seem an attractive explanation for N-terminal positive charge enrichment, other protein-structural origins for the use of positive charges should also be considered: just because positive charges slow ribosomes does not mean they have been selected to do so. A more architectural hypothesis might alternately envisage that the accumulation of positive charge at N-termini reflects some basic structural requirement of certain proteins. In this way of thinking, positive charge is not selected for because of its influence on a short-lived process (translation), but because of its contribution to the integral composition or structure of the protein itself. As positive charges have been well established to play a role in determining the orientation of integral membrane proteins, we here consider their usage as a possible alternative explanation for the N-terminal enrichment of positive charge.

The so-called "positive inside rule", which applies to proteins in both prokaryotes and eukaryotes, both with and without signal sequences, says that proteins orientate so that excess positive charge near hydrophobic membrane-spanning regions lies on the cytoplasmic side of the membrane (von Heijne and Gavel 1988; Sipos and von Heijne 1993). Correspondingly, the experimental addition of positively charged residues to normally periplasmic regions is capable of inverting the topology of a protein, such that the excess of positive charges will lie in the cytosol (Nilsson and von Heijne 1990). The insertion of proteins into membranes is thought to be achieved by a variety of conserved translocases and integrases (such as the well-described Sec translocon) acting both independently and cooperatively (Samuelson et al. 2000; Dalbey et al. 2011; Nishiyama et al. 2012). The addition of positive charges to the N-termini of transmembrane proteins can prevent the translocation of the termini across membranes in both *E. coli* and eukaryotes (Gafvelin et al. 1997), whether they require the main Sec protein-conducting channel (Li et al. 1988; Yamane and Mizushima 1988) or not (Whitley et al. 1994).

While the prevalence of the positive-inside rule is recognized, the mechanisms by which positive charges exert their topogenic effects are not well understood. Membrane protein topology may arise, at least in part, from positive charges near hydrophobic stretches stopping the transfer of further stretches of the protein across the membrane, thus anchoring the hydrophobic region within the bilayer (Kuroiwa et al. 1990). The positively charged residues might electrostatically interact with the negative phospholipid groups of the bilayer, preventing translocation of this portion of the protein through the membrane (Gallusser and Kuhn 1990; van Klompenburg et al. 1997). The proton-motive force leading to the acidification of the periplasm may be required for the translocation of some protein segments, facilitating transfer of negative but not positive residues across the membrane (Whitley et al. 1994; Kiefer et al. 1997; Delgado-Partin and Dalbey 1998). The arrangement of conserved positive and negative charges within the exoplasmic and cytoplasmic portions, respectively, of the Sec translocon itself could additionally contribute to the topogenesis of membrane proteins by interacting with charged residues within the proteins (Goder et al. 2004).

Can the positive inside rule alone explain the differential location of positively charged amino acids within proteins or do we in addition need to evoke selection on ribosomal velocity? The positive inside rule makes numerous predictions regarding which proteins and where in the proteins we expect to see positive charge enrichment. We test these predictions and show that the increase in average positive charge usage at the start of proteins is parsimoniously explained in full as a consequence of the need for many proteins to thread themselves through and orientate themselves in lipid bilayers. In both *E. coli* and *S. cerevisiae* we find increasing N-terminal charge amongst membrane proteins, not cytoplasmic ones. Focusing on *E. coli* (due to the need for large sample sizes of experimentally-supported transmembrane protein annotations), we find positive charge enriched at the point where cytosolically exposed N-termini enter the membrane, in accordance with the positive-inside rule. That similar patterns are repeated at the

C-terminus, where no ramping effect on downstream translation is to be expected, suggests the N-terminal positive charge pattern is purely protein-structural in origin. Cleavable signal sequences in *E. coli* tend to be rich in cations (von Heijne 1984) and lend an additional enrichment of positive charge at protein starts. We finally demonstrate that N-terminal positive charge patterns can be entirely explained by patterns of downstream cation usage in proximity to membranes. Thus the overall increase in positive charge across all N-termini results from the use of positive charge adjacent to the cytosolic leaflet of membranes combined with a skewed distribution of where cytosolic N-termini cross the plasma membrane.

Results

Across all three domains of life, average positive charge usage increases nearing protein N-termini

First we ask about the generality of the N-terminal loading of positively charged amino acids. While an increase, on the average, at N-termini of the density of positively charged amino acids has been seen in a few species, just how general is it? Upon aligning proteins by their N-terminus and calculating the average usage of positive charge within a given amino acid site, we observe that the average use of positive charge within 622 of 648 organisms (including the vast majority of Bacteria and Archaea studied) tends to increase nearing the N-terminus (Figure 1 and Figure S1). Given our constraints as regards which coding sequences we will include (see Methods, Sequences), we were only able to retain a small number of proteins for analysis for some eukaryotes, around 1-10% of the total number of genes encoded in the genome (see Figure S1). We consider it a strong possibility that the positive charge pattern is not seen in these organisms (Figure S1) due to the sequencing quality of these genomes. Indeed, it is only for such low-coverage eukaryotes that we do not observe significantly enriched N-terminal charge, quite possibly because we had to remove (via our sequence filters, see Methods) the subset of proteins that contribute to this pattern when all proteins are considered within an organism en masse.

This increasing use of positive charge near N-termini in 622 species is consistent with prior observations that mean charge increases nearing the N-terminus in *S. cerevisiae* and *E. coli* (Tuller et al. 2011). As we are interested in the potential ramifications of positive charge on ribosomal slowing, however, and we previously found no effect of negative charge on translation speed (Charneski and Hurst 2013), we consider only positive, not negative, charge here and in all further analyses.

Increasing N-terminal average positive charge is not found in non-membrane proteins

Investigating the positive charge pattern amongst groups of proteins that are differentially sublocalized within *E. coli* and *S. cerevisiae* shows that positive charge does not generally increase nearing N-termini in cytosolic proteins, but in proteins which are localized near to or potentially within membranes. In *E. coli*, increasing N-terminal charge is found amongst proteins generally localizing to the inner and outer membranes as well as periplasm, and in yeast, such a pattern is found in proteins resident near the mitochondrion, ER, Golgi, and vacuole (Figure 2; see Figure S2 for more yeast subcellular localizations). Hence the proteomic-scale pattern in *E. coli* and *S. cerevisiae* in Figure 1 results from the locations of positive charges in a subset of proteins within the organisms.

The increased positive charge at N-termini is associated with both the topology and sometimes signal sequences of transmembrane proteins

What is it about membrane proteins that leads to increasing N-terminal positive charge (Figure 2)? As the orientation of membrane proteins correlates with the density of positive charges on the cytoplasmic side of the membrane (von Heijne and Gavel 1988; Sipos and von Heijne 1993), we wondered whether the rise in positive charge at N-termini may be linked to the orientation of these termini. Indeed we find that among transmembrane proteins in both *E. coli* and *S. cerevisiae*, cytoplasmically-orientated N-termini show a far greater increase in positive charge at the tail than do those in the periplasm (Figure 3). We also note that cleavable signal sequences in *E. coli* tend to be positively charged (Figure 3A), as previously reported (von Heijne 1984). This is in line with findings that such charges in cleavable signal sequences, while not always essential for export, can significantly enhance the rate of translocation (Vlasuk et al. 1983; Puziss et al. 1989). This means that periplasmic proteins in *E. coli* display even more minimal N-terminal charge when proteins with N-terminal signal sequences are excluded (Figure 3A, last panel).

To determine if the enrichment of positive charge in transmembrane protein N-termini that are cytosolic is significantly different from those that are periplasmic, outside of any additional contribution of signal sequences, the following randomization was performed independently in *E. coli* and yeast. Signal-less proteins with N-termini in the cytosol and periplasm were combined into one group and then randomly sampled without replacement into two groups the same sizes as the observed groups. For each resampling, the average proportion of positive charge at each position in the first thirty amino acids was calculated in each of the two resampled groups. We then summed the differences in the average proportion of positive charge usage between the two sets at each N-terminal amino acid position, using a one-tailed approach as we have a strong prior that the cytoplasmic termini will display greater positive charge. If linear fits for the randomized N-cytosolic and N-periplasmic intersected before 30 amino acids downstream of the N-terminus, we stopped summing the differences at the point where the fits

intersected, otherwise we summed the differences over all 30 N-terminal amino acid positions. After 10,000 iterations, p was calculated as $(m + 1)/(n + 1)$, where n is the number of iterations and m is the number of times the randomized "area between the curves" was greater than or equal to that observed. This test indicates the chance of randomly obtaining such a large difference between the two curves in similar sized groups given the transmembrane proteins used to calculate those curves is rather low indeed in both *E. coli* ($p = 0.0001$) and yeast ($p = 0.0001$).

Thus we conclude the difference in positive charge usage between the two groups is highly significantly different, with positive charge loading occurring in the cytosolic N-termini of integral membrane proteins (in agreement with the positive-inside rule). This observation is straightforwardly interpreted in terms of the protein biochemistry/membrane orientation argument. In principle if one could propose a *post hoc* rationalization as to why membrane proteins in particular uniquely require the putative ribosomal slowing effects of positively charged residues, then this result can also be considered as not falsifying the positive-charge ramp model. We are unaware, however, of any such *post hoc* rationalization.

Positive charge is enriched in cytosolic N-tails near and just up to the point where the proteins enter the membrane

In the previous section we show that increased N-terminal positive charge is associated with an N-cytosolic transmembrane topology. We now look more closely at the configuration of cytoplasmic N-tails in relation to the plasma membrane, and investigate where positive charge tends to be used in relation to the point where these tails penetrate the membrane.

Although it would require a more specialized hypothesis to imagine a scenario in which only N-cytosolic membrane proteins require a positive charge driven ramp, we assume for the moment that such a hypothesis is possible for the sake of the argument. If positive charge is enriched in cytosolic N-termini (within the first 30 amino acids) to slow ribosomes, we might expect that within a given protein, positive charge tends to be used closer to N-termini than to the downstream region where the protein enters the membrane—or, perhaps we might expect no correlation at all between the densest regions of positive charge and their proximity to the membrane. However, if N-termini are enriched in positive charges to orientate proteins in membranes, we expect to see a bias in positive charge usage close to the point where the protein enters the membrane (von Heijne and Gavel 1988), with less positive charge usage upstream, closer to the initiating methionine. We examined the N-cytosolic regions of transmembrane proteins that were 10 to 30 amino acids in length and compared the density of positive charges in the first 5 amino acids at the N-terminus (following the first amino acid, normally an uncharged methionine)

to the density of positive charges in the five downstream cytosolic amino acids adjacent to the membrane. (The 10 amino acid minimum for these protein tails simply gives enough length to allow us to distinguish upstream, or N-terminal amino acids from downstream, membrane-adjacent ones). Not only do we find that 81% of proteins investigated have more positive charge in their membrane-adjacent region than upstream at the N-terminus (binomial test, $p < 2.2e-16$), but that the magnitude of positive charge in this membrane-adjacent region is significantly higher than in the upstream N-terminal region within the same protein (paired one-sided Wilcoxon test, $p < 2.2e-16$). Thus positively charged residues are used in proximity to the plasma membrane, not in proximity to the N-terminus per se.

We also find that the last positive charge used in an N-cytosolic segment tends to lie just near the face of the plasma membrane (Figure 4). We however find no such trend for positive charge usage for N-termini which lie on the periplasmic side of the bilayer (Figure 4). These findings are consistent with the above proposition that positive charge use at N-termini is linked to membrane proximity and the positive inside rule (Heijne 1986).

The degree of positive charge at the N-terminus corresponds to the length of transmembrane peptide exposed to the cytosol

That the increase in positive charge at cytosolic N-termini, that appears when averaged across membrane proteins, is actually a function of the point where the protein intersects the membrane and not a feature inherent to the N-terminus itself is well demonstrated visually. Upon progressively increasing the maximum length of N-cytosolic tails to be plotted, we see that the area of the N-terminus over which average positive charge increases is a function of the length of the exposed cytosolic tail (Figure 5A). This is in line with our finding that positive charges are used in the cytosolic portion of the protein before contacting the membrane. However, when we consider independent ranges of N-cytosolic lengths, it becomes apparent that positive charge is in fact not used more heavily in all proteins near the very beginning of proteins, but at the point where the protein meets the membrane (Figure 5B). Thus the positive charge curve for all N-cytosolic proteins (see Figure 3) appears to increase because the distribution of tail lengths is weighted towards the shorter end, with the majority of N-cytosolic tails being quite small (Figure S3). When combined with a tendency for higher positive charge usage in proximity to membranes, this length distribution creates the monotonic curve seen in Figure 3 (see Figure 6 for a graphical representation of this concept). This finding strongly argues for the protein orientation argument and against the ramp argument, as the ramp would propose (we presume) that all proteins should have the positive charges either randomly scattered or in approximately the same place.

Positive charge usage is also tied to transmembrane protein architecture at the C-terminus

We consider that if similar trends in positive charge use exist at C-termini, where no ramping effect on downstream translation should be expected, this would be strong evidence that N-terminal positive charge usage consequence of protein biochemistry rather than translational regulation. Indeed, we find that increasing positive charge usage nearing membrane protein C-termini is strong amongst those that lie in the cytosol, and remarkably minimal in those which are periplasmic (Figure S4). Amongst the transmembrane proteins which have between 10 and 30 amino acids in the cytosol at the C-terminus, more positive charge is found within the five amino acids closest to the membrane on the cytoplasmic side compared to the five most C-terminal amino acids (binomial test, $p = 0.016$), ignoring the last two amino acids of proteins since their basicity can greatly enhance translation termination efficiency (Mottagui-Tabar et al. 1994; Bjornsson et al. 1996). Additionally, this density of positive charge in the five amino acids just adjacent to the cytoplasmic face of the membrane is significantly greater than the magnitude of positive charge in the corresponding C-terminal region (paired one-sided Wilcoxon test, $p = 3.7e-05$). As might be expected if increasing positive charge usage at C-termini is tied to orientating proteins within membranes, the most upstream positively charged residue within the last thirty amino acids of a protein lies very close to the inner leaflet of the membrane (Figure S5). Similarly to the N-terminus, the degree of positive charge at the C-terminus is a function of the length of transmembrane tail which is exposed to the cytosol, with a combination of the lengths of C-tails exposed to the cytosol (Figure S3) and a tendency for positive charge to be used near membranes contributing to the emergent increasing charge pattern seen in C-cytosolic membrane protein termini (Figure S6).

Given these results, we conclude that transmembrane protein topology is capable of creating increasing average positive charge curves at either terminus, simply as a consequence of the membrane protein topologies at that terminus, without the need to invoke an N-terminal translational speed ramp to explain positive charge use at the beginnings of proteins.

Positive charge usage in proximity of transmembrane regions can entirely account for the increasing positive charge at the N-terminus

That similar phenomena contribute to increasing average positive charge nearing both N- and C-termini strongly suggests that N-terminal average positive charge patterns are caused by selection on transmembrane protein structures alone. As an additional control that no other major force is contributing to the increase in positive charge at N-termini, we asked whether N-terminal positive charge usage near membranes substantially differs from that observed in proximity to membrane-crossings that occur in the middle of the protein. Such transmembrane regions that lie further downstream in the

protein sequence allow us to measure membrane-proximal positive charge usage patterns outside any extra influence on positive charge usage at the N-terminus there might be.

We located all N-cytosolic into membrane transitions that occurred more than 45 amino acids downstream of the protein start. This allowed us to then create a profile of positive charge usage near these downstream membrane crossings. For each detected transition, we located six adjacent windows of five amino acids each: the first three windows lying in the cytoplasm, and the latter three windows in the membrane. We calculated the number of positive charges present in each window in each eligible protein, allowing us to eventually calculate the average density of positive charge in each downstream window position relative to the membrane. These average densities were then used to "reconstruct" the upstream (first thirty amino acids) N-terminal positive charge. For every protein which went into making Figure 3 (N-cytosolic proteins panel), the point at which the N-terminus hits the membrane was recorded, and the reconstructed positive charge for that protein was incremented in each possible five-amino acid window surrounding that point by the observed average density in that window position relative to the membrane. The observed N-terminal average positive charge pattern (Figure 3, N-terminus cytosolic) is not significantly different from this reconstructed N-terminal positive charge resulting from patterns of downstream positive charge usage patterns combined with the locations of where N-termini cross from the cytoplasm into membranes (Figure 7). Thus we infer that membrane protein topology alone is responsible for the increase in positive charge at the N-terminus. Our ability to reconstruct the increasing positive charge pattern in the "average protein" is an additional demonstration that such a pattern can and does indeed result from positive charge in transmembrane regions adjacent to the cytosolic leaflet in conjunction with a bias for short N-terminal lengths (Figure 6, Figure S3).

Do membrane proteins have more ribosomes on the N-termini?

The hypothesis that we were testing provided evidence that positive charges at N-termini correlated with increased ribosomal loading at transcript starts in yeast (Tuller et al. 2011). Do we then find increased ribosomal occupancy in membrane proteins as our findings might predict? Examining the relative change in ribosomal occupancy along a transcript (relative to the average occupancy of that transcript; see Methods), we find that ribosomal density is enriched, particularly over the first several codons, in transmembrane proteins compared to cytosolic ones (Figure 8). Another study similarly found an enrichment of ribosomal footprints in ER-associated proteins compared to cytosolic ones in a human embryonic kidney cell line (Reid and Nicchitta 2012). However, we observe that amongst membrane proteins, N-periplasmic proteins are either at least as or more enriched in ribosomal density along the first few codons of a transcript than N-cytosolic ones in both *E. coli* and *S. cerevisiae* (Figure 8), suggesting that increasing average positive charge density is not responsible for the most proximal

ribosomal densities (over the first few codons) observed on transcripts. After this initial excess, membrane proteins with cytosolic N-termini do appear to stay somewhat more occluded by ribosomes than other proteins, consistent with the result that positive charges correlate with average ribosomal density in yeast (Tuller et al. 2011). However, ramp-like densities are observed in all subclasses of protein (Figure 8), indicating some feature common to all subclasses is responsible for the bulk of the 5' ribosomal loading.

We have thus far assumed that the ramp observed in yeast is, like the increased positive charge at N-termini, also a phylogenetic universal. However, no such ramp was observed in mouse embryonic stems cells (Ingolia et al. 2011). What then about *E. coli*? Given increased usage of three features associated with ribosomal slowing (charge, codon bias and RNA folding) in the first 30+ codons of *E. coli* proteins, much as seen in yeast, it was presumed that *E. coli* would also have a ramp-like slowing effect (Tuller et al. 2011; see their Figure 1). However, scrutiny of Figure 8 indicates that, in contrast to yeast, apart from the ribosomal excess over the initial (first few) codons, there is no evidence for an extended ramp in *E. coli*, either in membrane proteins or cytoplasmic ones (Figure 8). Indeed after the initial ribosomal excess ($x > 4$) in *E. coli*, occupancy at each position is roughly the same as the average occupancy along the rest of the gene, and actually tends to increase somewhat from $x = 5$ to $x = 30$, counter to the above prediction (see Figure 8 caption).

Discussion

We find that the increasing use of positive charge nearing protein N-termini, seen when averaging over all proteins in a proteome, is due to transmembrane protein topology in both *E. coli* and *S. cerevisiae* (see Figure 6 for illustration). Such a finding is in accordance with positive charges being used to orientate N-tails in the cytosol as opposed to periplasm. The hypothesis that positive charge use at N-termini is due to membrane protein orientation makes correct predictions about which proteins have positively charged N-termini and where in proteins enrichment of positive charge is seen. While, on average, positive charge use increases monotonically approaching N-termini (Figures 1-3), in fact positive charge is used closer to membrane intersection point than to the N-terminus proper (Figures 4-5). Hence the overall increasing average positive charge pattern at protein starts is created by the length distribution of N-cytosolic tails of transmembrane proteins (Figure 5). That similar phenomena contribute to increasing average positive charge nearing C-termini strongly suggests that N-terminal charge patterns are simply a consequence of the structural needs of proteins, namely to orientate themselves in membranes in accordance with the positive-inside rule. That N-terminal average positive charge patterns can be entirely reconstituted from downstream positive charge usage patterns near membranes (Figure 7), where no selection on either ramping or other potential reasons for charge selection which might be particular to the N-terminus, further confirms the protein-structural basis for

this positive charge pattern. Importantly, we find no need to invoke a translational ramp to explain N-terminal positive charge densities.

Our results do not preclude that positive charges may be selected at termini for other physiochemical reasons. For example, cytoplasmically located proteins, while not displaying increasing charge nearing N-termini, do not show an absence of positive charge either (Figure 2). It is possible that within a subset of proteins (either transmembrane or cytosolic) an exposed tail may need positive charges to, amongst other things, bind other groups, including negative charges in nucleic acids (e.g. Moarefi et al. 2000), or that positive charge may be selected for at exposed termini residues to enhance protein solubility (e.g. Islam et al. 2012).

Our results, moreover, should not be over-interpreted. Our analysis is not designed to ask whether the ramp (as observed in yeast) is real or whether the ramp is adaptive. We simply wish to know whether the increase in average positive charge nearing N-termini, which we have shown to be a widespread, if not phylogenetically universal pattern, is best explained as part of a mechanism to stall ribosomes. We cannot on the basis of our results conclude that there is no ramp in any organism, or that any potential ramp is necessarily not adaptive. We note, rather, that the loading of positive charges at N-termini is not evidence that such stalling is adaptive or that positive charges are selected at N-termini for gene regulatory purposes, especially as we find a more parsimonious explanation for the presence of these charges. Consistent with positive charge being better explained by a factor other than a regulatory ramp, we find no evidence for a 5' ribosomal ramp of the predicted dimensions (at least 31 codons) in any class of protein upon examination of ribosomal footprinting data in *E. coli* (Figure 8). This is a surprising result as 5' aberrations in codon usage (Eyre-Walker and Bulmer 1993), encoded positive charge (Berezovsky et al. 1999), mRNA folding (McCarthy and Bokelmann 1988; de Smit and van Duin 1990), or a combination of the three were predicted to cause a 5' increase ribosomal densities along a transcript (Tuller et al. 2011).

Taken together with other recent results, however, our findings add a little to the literature questioning the validity of the adaptive ramp hypothesis. Some have questioned, as we did, whether features of the ramp are better explained in other terms. The hypothesized ramp is posited to be, in addition to a consequence of charge, caused by two other properties of 5’ ends of mRNAs: non-optimal codon usage and strong RNA folding (Tuller et al. 2010). Leaving aside the problem that analysis of ribosome protection data failed to find evidence that non-optimal codons slow ribosomes under normal conditions (Qian et al. 2012; Charneski and Hurst 2013), there is an alternative and more parsimonious interpretation of the enrichment of rare codons at 5’ ends, in terms of reducing (not increasing) RNA folding stability to enable translation initiation (Bentele et al. 2013). Combined with our analysis, the

inference that rare codons and positive charges are enriched at 5' ends/N-termini to enable ribosome slowing now seems an unparsimonious model.

An immediate problem for the ramp hypothesis is our observation that any excess ribosomal occupancy is seen only at the very start of transcripts in *E. coli.* Why might this be? The ramp is defined as a net increase in mean ribosomal occupancy as one moves towards the 5' end of transcripts. Recently, it has been suggested that high initiation rates on shorter transcripts will give a higher mean occupancy at 5' ends when averaged over multiple transcripts of all lengths, but need not necessarily be seen in any given transcript (Shah et al. 2013). In principle, if initiation rates are not biased towards small transcripts in *E. coli*, such a statistical artifact could explain the apparent species differences (Figure 8). However, our analysis is normalized by mean transcript occupancy (Figure 8). As we see on average a downward trend in yeast, a factor other than, or in addition to, short transcripts undergoing more frequent initiation may be required to explain all of the observed 5' ribosomal densities in this organism.

A further issue for the ramp hypothesis is whether the high ribosomal occupancy seen in both *E. coli* and yeast in close proximity to the start codon (~4 codons in *E. coli*, ~6 in yeast; Figure 8) need reflect elongating ribosomes, as the ramp model presumes. In both eukaryotes and prokaryotes it is possible for small subunits which have not yet bound a large subunit or completed initiation to bind the start site (Kozak 1999). It may be possible that such non-initiated small subunits are detected by general endonuclease footprinting protocols. Indeed similarly to our results (Figure 8), other footprinting datasets in *E. coli* (Oh et al. 2011) and a human embryonic kidney cell line (Reid and Nicchitta 2012) also profiled short spikes in ribosomal density over only a few codons at the most 5' ends of transcripts. These short occluded distances at the extreme 5' end are roughly consistent with the 16-17 nt which are occluded at the start of the coding sequence by a small subunit at the start codon in yeast (Anthony and Merrick 1992). We suggest that the differences in elevated average relative ribosomal densities (along the first few codons at least) in all plotted protein subgroups may to some extent simply reflect differences in translation initiation rates.

The above interpretation may be consistent with apparently different results, dependent on method (Ingolia et al. 2011). For example, addition of the non-hydrolyzable GTP analog GMP-PNP prevents full (GTP-dependent) ribosome initiation complex formation and leads to an accumulation of small ribosomal subunits positioned at start codons, occluding about 16-17 nt at the start of the coding sequence (Anthony and Merrick 1992). The use of an initiation inhibitor in making the *E. coli* dataset under consideration (Li et al. 2012) could then potentially exacerbate the problem of enriched footprints in this region that in fact correspond to non-elongating ribosomes. Additionally, the 5' charge density may also arise from the fact that GMP-PNP should not have an effect on already-formed initiation complexes that are ready to immediately start translating. Such pre-formed complexes might be able to

translate a couple of codons before the elongation inhibitor chloramphenicol (Li et al. 2012) or cyclohexamide (Ingolia et al. 2009) are able to act. Such a mechanism is consistent with the slight upswing in ribosomal density in the 1-2 codons in *E. coli* or 4 codons in *S. cerevisiae* after the translational start (Figure 8). It is also consistent with the wider initial ribosomal excess (over ~6 codons) in yeast compared to *E. coli*, which might result from fewer codons being strictly stalled over the start codon by the initiation inhibitor.

We must emphasize that we do not wish to claim it is necessarily the case that at least some of the increased ribosomal density at 5' transcript ends is an artifact of the methods used to suppress translation, merely that it is a possibility. There might be true taxon-specific differences in initiation or elongation that cause the observed differences in 5' ribosomal densities in either *E. coli* or yeast, or both. Nor has the potentially confounding issue of ribosomal drop off yet been addressed. Understanding to what extent a 5' excess of ribosomes is a result of true taxonomic differences versus a methodological and/or statistical artifact must be a high priority. It remains to be discovered whether any transcript or protein feature has been under selection to modulate ribosome velocity.

Methods

Sequences

The June 2008 release of Saccharomyces Genome Database gene sequences was obtained from the eukaryotic UCSC Table Browser (Karolchik et al. 2004) at http://genome.ucsc.edu/ and most other eukaryotic coding sequences were obtained from http://genome.ucsc.edu/ on 6 April 2013. However, in order to facilitate analysis of ribosomal footprinting data by Ingolia et al., annotations of the *S. cerevisiae* S288C genome as available on June 22, 2008 (the build used by Ingolia et al. 2009) were obtained separately from the *Saccharomyces cerevisiae* Genome Database (www.yeastgenome.org); only protein-coding sequences of non-dubious classification were considered. Archaeal RefSeq (and the bacterial Refseqs used for making Figure 1) nucleotide sequences coding for protein were downloaded from the microbial UCSC table browser via http://microbes.ucsc.edu/ on 26 March 2011. For bacterial genomes we considered the latest release of the EMBL bacterial genome set (http://www.ebi.ac.uk/genomes/bacteria.txt), downloading each genome from EBI using a purpose written web-crawler. We then considered one genome from each bacterial genus.

For all organisms, any sequences containing nonsense codons or which were not multiples of three were excluded. The sequences were further filtered to only allow the standard or alternative start codons indicated in the appropriate NCBI genetic code table from http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi. Table 1 was used for Eukaryotes, Table 4

for mycoplasmas, Table 11 for other Bacteria as well as Archaea. The remaining nucleic acid sequences were translated into protein according to these tables.

Protein localizations

E. coli cytoplasmic (and other) proteins for Figure 2 were obtained from Table S5 of Han et al. (2011). Only those proteins which had all forms of evidence supporting their localization were considered. While another attempt to sublocalize *E. coli* proteins has been made (Lopez-Campistrous et al. 2005), this dataset is much smaller, and there is little overlap in the Swissprot annotations to which the authors compare their localizations, with only 42 proteins agreed to be in the cytosol by both sources. For these reasons we use the larger dataset (Han et al. 2011).

S. cerevisiae localizations experimentally determined using a GFP reporter construct (Huh et al. 2003) were downloaded from http://yeastgfp.yeastgenome.org. Proteins were allowed to localize to more than one location.

Protein topology

The current release of TOPDB, a membrane protein topology database which is based on experimental structural and topological information (Tusnady et al. 2008), was downloaded in xml format from http://topdb.enzim.hu on 25 March 2013. We limit our analyses using information from this database to *E. coli* due to need of a sufficient sample size of transmembrane proteins within an organism. N-periplasmic peptide signals were taken from the relevant TOPDB annotations.

Membrane protein topologies based on experimental protein fusions for *S. cerevisiae* were taken from Supporting Table 2 of Kim et al. (2006). All C-termini in this table were incorporated in our analysis as they all have direct experimental evidence supporting their topology. Those N-termini with topologies supported by both HMM models were considered in the main text. Yeast proteins with signal peptides were downloaded from the Ensembl Biomart dataset EF4 at http://www.ensembl.org/biomart/.

Calculating the average proportion of positive charge in a given site across proteins

To calculate the tendency for positive charge to be used at a certain distance from the N- or C-terminus within a given species, the set of proteins under consideration were aligned by their N- or C-, as appropriate, termini. The amino acids arginine, lysine and histidine were assigned a charge of 1 and all other amino acids were assigned as 0 (not positively charged). The average proportion of positive charge was then calculated in aligned positions. In all analyses, the first amino acid at the N-terminus is ignored

since it is always uncharged. If a protein is less than 60 amino acids in length, only half of the residues within that protein were considered in order to prevent interference of selection on charge at the opposite terminus. All plots consider the protein terminus (C- or N-, as appropriate) to be at $x = 0$.

Determination of increasing average positive charge at N-termini

Some plots of the average proportion of positive charge along the aligned 30 most N-terminal amino acids are best fitted by higher order equations (as determined by ANOVA of nested models). To provide a statistic for whether average positive charge usage increases nearing N-termini in these cases, we note that for an equation of the form $y = ax^n + bx + c$, the slope at any point on the curve is given by $dy/dx = n \oplus ax^{n-1} + b$. Thus at the extreme ($x = 0$), regardless of the order of the regression, $dy/dx = b$. This means if the linear term coefficient b is negative, we infer the use of positive charge increases approaching the N-terminus, and the strength of the increase is reflected in the magnitude of b. Linear models and other statistical analyses generally were done in R (R Development Core Team 2010).

Robustness of increasing or decreasing positive charge patterns at termini to topology annotation

Our inference that the positive charge usage at N-termini results from the cytosolic orientation of the N-terminus relies upon the presumption that the TOPDB database we use does not rely on the use of positive charge to assign topologies. To this we note the annotations in the TOPDB database are based on experimental evidence, both structural and topological (Tusnady et al. 2008). In combination with this information, the database uses an HMM algorithm (Tusnady and Simon 1998) trained on an experimentally-determined, well-defined set of topologies to help predict unknown topologies. While the HMM considers that different structural parts of a protein (e.g. transmembrane segments, loops) are likely to show an amino acid composition which is divergent compared to the amino acid usage of the protein as a whole, it makes no stipulations about what those amino acid compositions must be. That is, the HMM does not assign membrane topologies by enforcing predetermined rules governing the usage of positive charge, or any other physiochemical property of amino acids, on either side of a transmembrane region.

Nonetheless, we wanted to ensure that the increasing positive charge pattern we detect at cytosolic N-termini is not the result of positive charges in the N-termini of the training set being propagated (or erroneously propagated) through to the topology prediction of N-termini. We find our results in *E. coli* are indeed robust to using topology annotations supported by increasing levels of experimental evidence, including experimental evidence gathered at the N-terminus specifically (Figure S7).

We also note that all the HMM-predicted N-termini topologies in yeast (Kim et al. 2006) are constrained by experimental information regarding the topology of the C-terminus that the authors produced in the same paper. One of the HMMs used to predict the N-terminal topologies, prodiv-TMHMM, relies on a similar method to the one used by TOPDB, and does not explicitly use positive charge to infer the most likely membrane orientation (Viklund and Elofsson 2004). The other HMM employed by Kim et al., TMHMM, does incorporate (amongst other factors) charge bias in its determinations of membrane protein topology (Krogh et al. 2001). In the main text we use those proteins whose topologies are supported by both of these two independent methods. For completeness and transparency we have also examined the increase in positive charge use amongst those proteins topologies predicted by each HMM separately. We can report that these additional analyses give similar results to those presented in the main text, namely that increasing N-terminal positive charge is observed only amongst those membrane proteins whose N-termini reside in the cytosol (and in the absence of signal sequences) (Figure S8).

Determination of positions of significant average positive charge enrichment at N-termini

Within a group of proteins for which we perform a regression of positive charge usage on distance from N-terminus, we determined the location(s) of significantly increased average positive charge by 1,000 iterations of the following method. The sequences of all proteins within the considered group were shuffled and the proportion of positive charge within the randomized sequences was calculated in each of the first 30 positions near the N-terminus. For each iteration, if the proportion of randomized positive charge in a given position is greater than or equal to the proportion of positive charge observed in that position within the considered group, m is incremented in that position. P for each position is then calculated as $(m + 1)/(n + 1)$, where n is the number of iterations performed.

Determination of the point of maximal or minimal positive charge usage

For second-order equations and higher, the point of maximal or minimal charge usage corresponds to the point where the slope of the tangent is zero. This point was determined by setting the derivative of each linear model equal to zero and solving the equations in MATLAB (MATLAB 2010).

Reconstruction of N-terminal positive charge according to trends in charge usage near transmembrane regions

If proteins transitioning from the cytosol into membranes can indeed account for the increasing positive charge usage at the N-terminus, we should be able to reproduce the observed pattern of increasing average positive charge approaching cytosolic N-termini given solely the locations of where these

cytosol-to-membrane transitions occur. In order to measure trends in positive charge usage near membranes outside of any possible additional selection on positive charge within the first 30 amino acids, we consider in this analysis only those transmembrane regions within proteins where the cytosolic N-terminus transitions into the membrane at least 45 amino acids downstream of the start of the protein. For each protein with such a region, we recorded the number of positive charges used in each of six consecutive windows, each five amino acids in length, directly surrounding the cytosolic face of the membrane such that the first three windows cover cytosolic amino acids and the latter three windows cover amino acids situated in the membrane. The average number of positive charges in each window (defined by its location relative to the membrane) was then calculated across all suitable proteins.

We then returned to the distribution of locations where cytoplasmic N-termini come into contact with membranes within the first 30 amino acids at the N-terminus. The positive charge at each of (up to) 30 amino acid positions surrounding the N-terminal point of transition into the membrane was incremented by the density of positive charges within the analogous observed window as calculated above. The average number of positive charges, as reconstructed, at each position was then calculated.

Ribosomal footprint data

Ribosomal densities derived from two replicates of ribosomally protected fragments along the transcriptome of *E. coli* (Li et al. 2012) were downloaded from GSM872393 and GSM872394 at http://www.ncbi.nlm.nih.gov/geo/. Positions in each replicate that had no footprint counts available were given a footprint count of zero. The ribosomal footprint counts at each position along the transcriptome were averaged between the two replicates.

Sequenced ribosomally protected fragments for *S. cerevisiae* grown in rich media, dataset GSE13750 (Ingolia et al. 2009) were downloaded from the NCBI Gene Expression Omnibus at www.ncbi.nlm.nih.gov/projects/geo. Only one mismatch between the sequenced fragment and reference genome sequence was allowed. The chromosomal location and coordinates of the sequenced fragments given in the original dataset were used in combination with the start and stop coordinates from the gene annotations (see Methods, Sequences) to map footprints to transcripts. All fragment counts were taken as the average value of the two experimental replicates and only footprints that mapped uniquely to one location in the reference genome were considered. In line with Ingolia et al., we assigned footprints to genes if the first base of the footprint mapped to 16 nt before the first base or 14 nt before the last base of the gene, in order to take account of which area of the footprint is likely in the ribosomal active site.

We consider that whether ribosomal occupancy is, on average, enriched at transcript starts is best addressed by normalizing the ribosomal density at the start of a given transcript relative to the average

ribosomal occupancy of *the same transcript*. This has the advantage of treating footprints as increases or decreases in density within the context of a single transcript (a ribosome travels only along a single mRNA at a time) and sidesteps the ambiguity in deciphering emergent patterns that might result from raw footprint counts being averaged across different transcripts. To this end, we calculated relative ribosomal occupancies for the first thirty codons ($1 \geq x \leq 30$) in each *E. coli* transcript by dividing the ribosomal density at position x by the average ribosomal occupancy of the entire gene. Transcripts were then aligned by their 5' ends and the mean relative ribosomal occupancy in each position was calculated to create Figure 8.

Funding

This work was supported by a University of Bath Overseas Research Studentship (to C.A.C.) and a Wolfson Royal Society Research Merit Award (to L.D.H.).

References

Anthony DD, Merrick WC. 1992. Analysis of 40 S and 80 S complexes with mRNA as measured by sucrose density gradients and primer extension inhibition. J Biol Chem 267: 1554-1562.

Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol 9: 675.

Berezovsky IN, Kilosanidze GT, Tumanyan VG, Kisselev LL. 1999. Amino acid composition of protein termini are biased in different manners. Protein Eng 12: 23-30.

Bjornsson A, Mottagui-Tabar S, Isaksson LA. 1996. Structure of the C-terminal end of the nascent peptide influences translation termination. EMBO J 15: 1696-1704.

Charneski CA, Hurst LD. 2013. Positively charged residues are the primary determinants of ribosomal velocity. PloS Biol 11: e1001508.

Dalbey RE, Wang P, Kuhn A. 2011. Assembly of bacterial inner membrane proteins. Annu Rev Biochem 80: 161-187.

de Smit MH, van Duin J. 1990. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc Natl Acad Sci U S A 87: 7668-7672.

Delgado-Partin VM, Dalbey RE. 1998. The proton motive force, acting on acidic residues, promotes translocation of amino-terminal domains of membrane proteins when the hydrophobicity of the translocation signal is low. J Biol Chem 273: 9927-9934.

Dimitrova LN, Kuroha K, Tatematsu T, Inada T. 2009. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. J Biol Chem 284: 10343-10352.

Eyre-Walker A, Bulmer M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Research 21: 4599-4603.

Gafvelin G, Sakaguchi M, Andersson H, von Heijne G. 1997. Topological rules for membrane protein assembly in eukaryotic cells. J Biol Chem 272: 6119-6127.

Gallusser A, Kuhn A. 1990. Initial steps in protein membrane insertion. Bacteriophage M13 procoat protein binds to the membrane surface by electrostatic interaction. EMBO J 9: 2723-2729.

Goder V, Junne T, Spiess M. 2004. Sec61p contributes to signal sequence orientation according to the positive-inside rule. Mol Biol Cell 15: 1470-1478.

Han MJ, Yun H, Lee JW, Lee YH, Lee SY, Yoo JS, Kim JY, Kim JF, Hur CG. 2011. Genome-wide identification of the subcellular localization of the Escherichia coli B proteome using experimental and computational methods. Proteomics 11: 1213-1227.

Heijne G. 1986. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 5: 3021-3027.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. Nature 425: 686-691.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218-223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789-802.

Islam MM, Khan MA, Kuroda Y. 2012. Analysis of amino acid contributions to protein solubility using short peptide tags fused to a simplified BPTI variant. Biochim Biophys Acta 1824: 1144-1150.

Ito-Harashima S, Kuroha K, Tatematsu T, Inada T. 2007. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. Genes Dev 21: 519-524.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: D493-496.

Kiefer D, Hu X, Dalbey R, Kuhn A. 1997. Negatively charged amino acid residues play an active role in orienting the Sec-independent Pf3 coat protein in the Escherichia coli inner membrane. EMBO J 16: 2197-2204.

Kim H, Melen K, Osterberg M, von Heijne G. 2006. A global topology map of the Saccharomyces cerevisiae membrane proteome. Proc Natl Acad Sci U S A 103: 11142-11147.

Kozak M. 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234: 187-208.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.

Kuroiwa T, Sakaguchi M, Mihara K, Omura T. 1990. Structural requirements for interruption of protein translocation across rough endoplasmic reticulum membrane. J Biochem 108: 829-834.

Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538-541.

Li P, Beckwith J, Inouye H. 1988. Alteration of the amino terminus of the mature sequence of a periplasmic protein can severely affect protein export in Escherichia coli. Proc Natl Acad Sci U S A 85: 7685-7689.

Lopez-Campistrous A, Semchuk P, Burke L, Palmer-Stone T, Brokx SJ, Broderick G, Bottorff D, Bolch S, Weiner JH, Ellison MJ. 2005. Localization, annotation, and comparison of the Escherichia coli K-12 proteome under two states of growth. Mol Cell Proteomics 4: 1205-1209.

Lu J, Deutsch C. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. J Mol Biol 384: 73-86.

Lu J, Kobertz WR, Deutsch C. 2007. Mapping the electrostatic potential within the ribosomal exit tunnel. J Mol Biol 371: 1378-1391.

MATLAB. 2010. Version 7.11.0 (R2010b). Version 7.11.0 (R2010b). Natick, Massachusetts: The MathWorks Inc.

McCarthy JE, Bokelmann C. 1988. Determinants of translational initiation efficiency in the atp operon of Escherichia coli. Mol Microbiol 2: 455-465.

Moarefi I, Jeruzalmi D, Turner J, O'Donnell M, Kuriyan J. 2000. Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. J Mol Biol 296: 1215-1223.

Mottagui-Tabar S, Bjornsson A, Isaksson LA. 1994. The second to last amino acid in the nascent peptide as a codon context determinant. EMBO J 13: 249-257.

Nilsson I, von Heijne G. 1990. Fine-tuning the topology of a polytopic membrane protein: role of positively and negatively charged amino acids. Cell 62: 1135-1141.

Nishiyama K, Maeda M, Yanagisawa K, Nagase R, Komura H, Iwashita T, Yamagaki T, Kusumoto S, Tokuda H, Shimamoto K. 2012. MPIase is a glycolipozyme essential for membrane protein integration. Nat Commun 3: 1260.

Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, Nichols RJ, Typas A, Gross CA, Kramer G, Weissman JS, Bukau B. 2011. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell 147: 1295-1308.

Puziss JW, Fikes JD, Bassford PJ, Jr. 1989. Analysis of mutational alterations in the hydrophilic segment of the maltose-binding protein signal peptide. J Bacteriol 171: 2303-2311.

Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J. 2012. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLoS Genet 8: e1002603.

R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Reid DW, Nicchitta CV. 2012. Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. J Biol Chem 287: 5518-5527.

Samuelson JC, Chen M, Jiang F, Moller I, Wiedmann M, Kuhn A, Phillips GJ, Dalbey RE. 2000. YidC mediates membrane protein insertion in bacteria. Nature 406: 637-641.

Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. Cell 153: 1589-1601.

Sipos L, von Heijne G. 1993. Predicting the topology of eukaryotic membrane proteins. Eur J Biochem 213: 1333-1340.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141: 344-354.

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12: R110.

Tusnady GE, Kalmar L, Simon I. 2008. TOPDB: topology data bank of transmembrane proteins. Nucleic Acids Res 36: D234-239.

Tusnady GE, Simon I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283: 489-506.

van Klompenburg W, Nilsson I, von Heijne G, de Kruijff B. 1997. Anionic phospholipids are determinants of membrane protein topology. EMBO J 16: 4261-4266.

Viklund H, Elofsson A. 2004. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci 13: 1908-1917.

Vlasuk GP, Inouye S, Ito H, Itakura K, Inouye M. 1983. Effects of the complete removal of basic amino acid residues from the signal peptide on secretion of lipoprotein in Escherichia coli. J Biol Chem 258: 7141-7148.

von Heijne G. 1984. Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells. EMBO J 3: 2315-2318.

von Heijne G, Gavel Y. 1988. Topogenic signals in integral membrane proteins. Eur J Biochem 174: 671-678.

Whitley P, Zander T, Ehrmann M, Haardt M, Bremer E, von Heijne G. 1994. Sec-independent translocation of a 100-residue periplasmic N-terminal tail in the E. coli inner membrane protein proW. EMBO J 13: 4653-4661.

Yamane K, Mizushima S. 1988. Introduction of basic amino acid residues after the signal peptide inhibits protein translocation across the cytoplasmic membrane of Escherichia coli. Relation to the orientation of membrane proteins. J Biol Chem 263: 19690-19696.

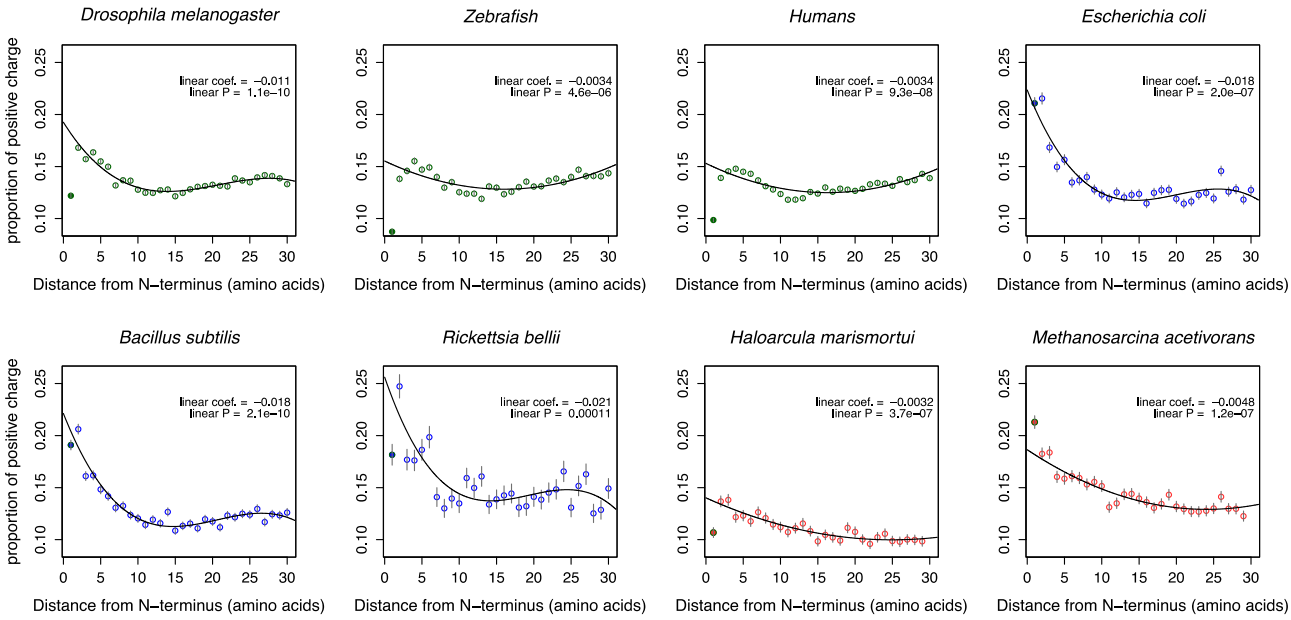

Figure 1. Average positive charge usage in an organism increases towards the N-terminus across all three domains of life. Whether the use of positive charge increases nearing the N-terminus is determined by the sign of the linear coefficient term (see Methods). Representative genomes are shown from Eukaryotes (green), Bacteria (blue), Archaea (red). See Figure S1 for more plots from 648 species. Here and in all plots error bars represent the standard error of the mean. The first amino acid following the initiating methionine (filled points) is often an outlier and was ignored in all regressions.

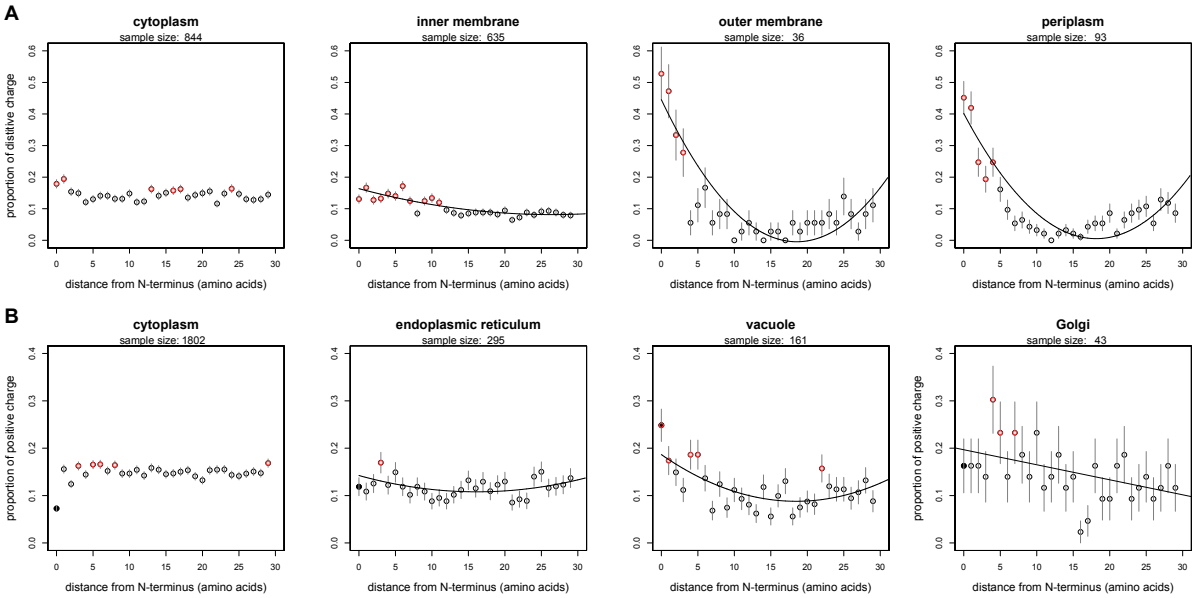

Figure 2. Only proteins potentially associated with membranes, not cytosolic proteins, show increasing N-terminal charge. Note that locations in this Figure correspond to general subcellular locations of entire proteins, but say nothing specifically about the exact locations of protein termini, which in the case of transmembrane proteins may vary depending on their orientation in the membrane. In addition, proteins classified as "membrane" proteins may for example be only peripherally bound (Han et al. 2011). Red points show positions of significantly enriched ($p < 0.05$) positive charge (see Methods). Linear versus quadratic best fits were determined by ANOVA of nested models. Whether y is increasing approaching the end of the protein ($x = 0$), in the quadratic regressions, is determined by the sign of the x-term coefficient (see Methods). **A. *E. coli.*** Cytosol: no increase in charge at N-termini: regression of $y \sim x$, slope $p = 0.20$. Outer membrane and periplasm: regression of $y \sim x^2+x$, fitted x-term value of -0.049 (p = 5.4e-09), $r^2 = 0.74$; -0.044 (p = 1.15e-10), $r^2 = 0.80$, respectively. Inner membrane: regression of $y \sim x^2+x$, fitted x-term value of -0.00622 ($p = 0.00017$), $r^2 = 0.68$. **B. *S. cerevisiae*.** Cytosol: no increase in charge at N-termini: regression of $y \sim x$, slope $p = 0.66$. Endoplasmic reticulum and vacuole, regression of of $y \sim x^2+x$, fitted x-term value of -0.0043 ($p = 0.033$), $r^2 = 0.10$; -1.1e-02 ($p = 0.0012$), $r^2 = 0.32$, respectively. Golgi, regression of $y \sim x$, linear coefficient -0.0032, $p = 0.010$, $r^2 = 0.19$. For more subcellular locations in yeast see Figure S2.

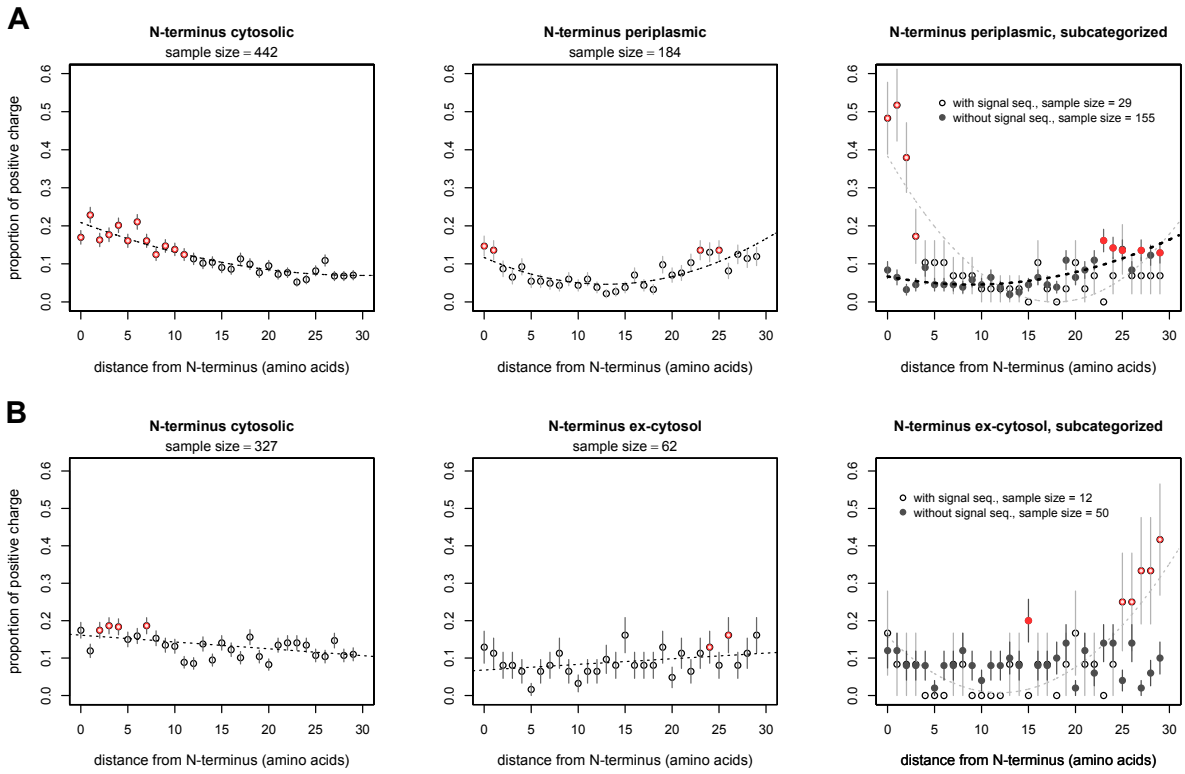

Figure 3. The topology and signal sequences of transmembrane proteins cause N-terminal positive charge loading. All proteins considered in these plots are transmembrane. Linear versus quadratic fits were determined by ANOVA of nested models. Whether y is increasing approaching the end of the protein (x = 0) any order of regression is determined by the sign of the linear term coefficient (see Methods). Red points show positions of significantly enriched ($p < 0.05$) positive charge (see Methods). Rows A and B have different y-axes. **A. *E. coli.*** All regressions are of the form $y \sim x^2+x$. *N-termini in the cytosol*: with signal sequences, x-term value of -0.0093 (p = 5.0e-06), r^2 = 0.82. There are no proteins with cytosolic N-termini which have signal sequences. *N-termini in the periplasm*: all proteins, fitted x-term value of -0.011 (p = 6.6e-06), r^2 = 0.61; only those with signal sequences, x-term -0.043 (p = 6.8e-08), r^2 = 0.69; those without signal sequences, x-term coefficient -0.0048 (p = 2.5e-02), r^2 = 0.62. **B. *S. cerevisiae.*** *N-termini in the cytosol*: $y \sim x$ slope -0.0018 (p = 0.0029), r^2 = 0.25. *N-termini ex-cytosol*: $y \sim x$ slope 0.0015 (p = 0.045), r^2 = 0.10; only those with signal sequences, $y \sim x^2+x$ regression x-term -0.025 (p = 2.6e-05), r^2 = 0.74; those without signal sequences, $y \sim x$ slope p = 0.94.

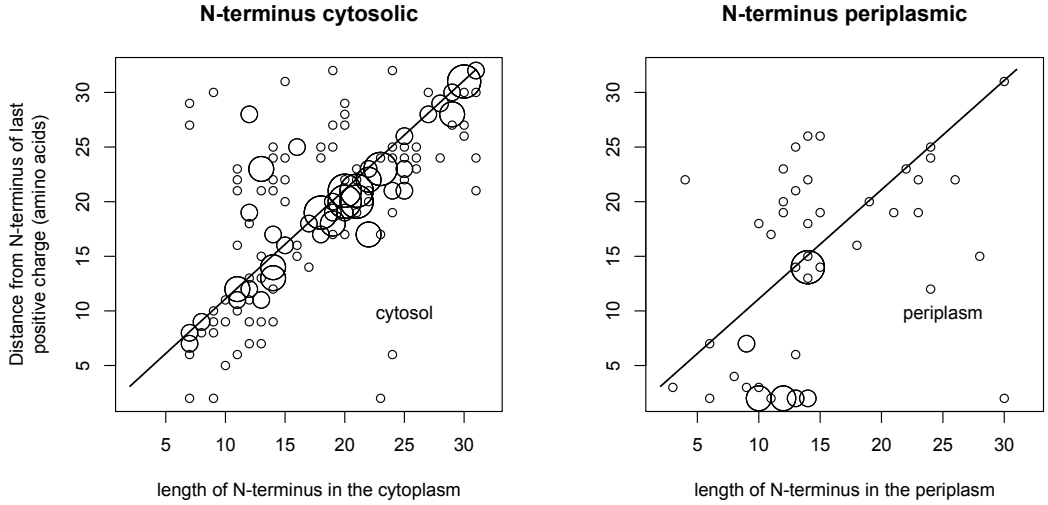

Figure 4. Among *E. coli* transmembrane proteins, the last positively charged amino acid of cytoplasmic N-termini tends to lie near the inner leaflet of the membrane. The size of the point is proportional to the number of times that point is plotted. The length of the N-terminal fragment must be less than or equal to 30 residues, purely because this is the length of the major phenomenon we are trying to investigate (see Figure 3A, N-cytosolic proteins). In either plot, if the use of positive charge is closely associated with membranes we should expect dense points near the line $x = y$ which represents the *face* of the appropriate membrane (the membrane itself will occupy more space above the thin line depicted). We note that the TOPDB protein topologies used in making this Figure are supported by experimental evidence and hence the trends we report here are not an artefact of prediction algorithms (see also Methods). ***N-terminus cytosolic***. The point of the next membrane crossing—i.e. where the N-terminus exists the membrane into the periplasm—must occur at least 31 residues downstream of the start of the protein, so as to not interfere with the N-terminus-into-cytosolic leaflet transition we wish to inspect. The horizontal line represents the inner face of the inner membrane and is depicted for visual purposes only. Spearman's rho between x and y, 0.67, $p < 2.2e\text{-}16$; the slope of a standardized major axis regression of $y \sim x$ is not significantly different from 1 ($p = 0.18$; slope coefficient 95% CI: 0.97, 1.2). Binomial test that positive charges have a 50/50 chance of being found on either side of the inner leaflet of the inner membrane, $p < 2.2e\text{-}16$ (with 156 out of 194 observations located leading up to and just at the cytosolic side of the membrane). ***N-terminus periplasmic***. Proteins with signal sequences are excluded from the plot as we wish to investigate the remaining interaction of the protein with the membrane once they are cleaved. Similarly to above, the point where the N-terminus exists the membrane into the cytoplasm must occur at least 31 residues downstream of the start of the protein. The horizontal line represents the outer face of the inner membrane and is depicted for visual purposes only. Spearman's rho between x and y, 0.47, $p = < 0.00033$; the slope of a standardized major axis regression of $y \sim x$ is significantly different from 1 ($p = 0.0057$; slope coefficient 95% CI: 1.1, 1.8). Binomial test that positive charges have a 50/50 chance of being found on either side of the inner leaflet of the inner membrane, $p = 0.00018$ (with 41 out of 54 observations on the periplasmic side of the membrane).

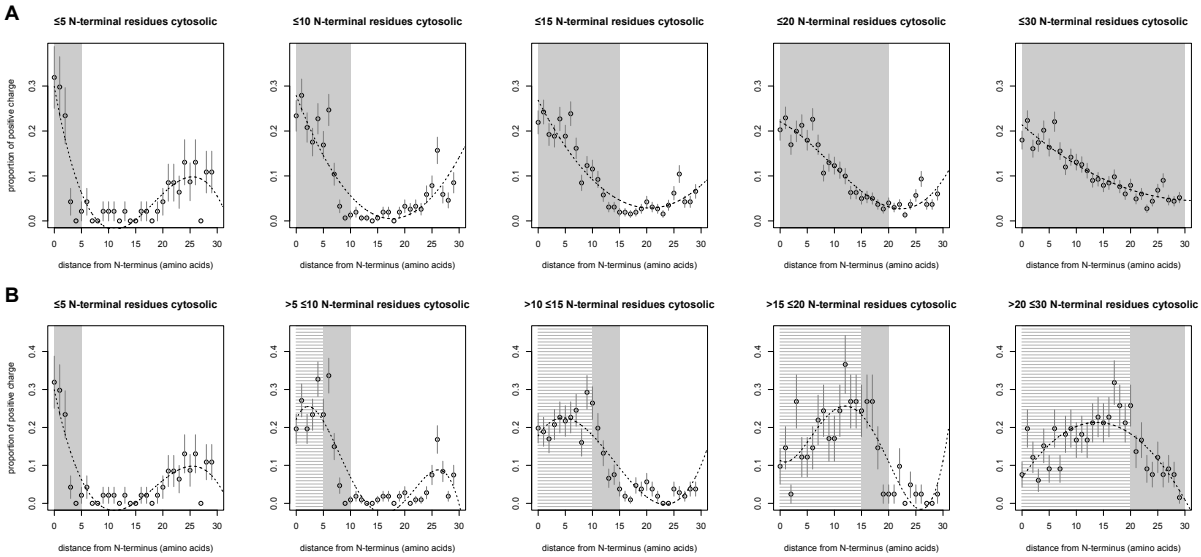

Figure 5. The degree of positive charge at the N-terminus corresponds to the length of transmembrane peptide exposed to the cytosol. All data is from *E. coli* and all proteins considered in plots are transmembrane. Higher order best fits were determined by ANOVA of nested models. The first panels in rows A and B are the same. **A**. Proteins with a maximal length of the N-terminus within the cytosol are considered, hence all subsequent plots encompass data from the previous plots within this row. The solid shaded region on each plot represents the possible range of locations wherein all the N-cytosolic proteins considered in that plot must cross into the plasma membrane. In other words, all the cytosolic portions in these plots reside within the solid shaded regions, but the shaded regions may also contain non-cytosolic residues if the protein crosses into the inner membrane before the shaded region ends. Average positive charge usage is well predicted by the propensity for the N-terminus to be cytosolic, as shown by a regression of *point of inflection* on *x*-axis where positive charge usage is lowest ~ *no. cytosolic N-terminal residues*: slope = 0.79, slope $p = 0.002$, $r^2 = 0.96$. **B**. Each plot considers a range of residue lengths that are exposed to the cytoplasm at the N-terminus, i.e. the plots in this row consider mutually exclusive sets of proteins. The solid shaded region, as in row A, represents the range of locations where the N-cytosolic protein transitions into a membrane. The region with striped, lighter shading represents the range of residues which can be guaranteed to reside within the cytosol. When considered this way, it becomes apparent that positive charge is used, on average, more and more just up to the point where the protein meets the inner face of the (negatively charged) phospholipid bilayer: regression of *point of maximal positive charge usage* along *x*-axis ~ *minimum no. of cytosolic N-terminal residues:* slope = 0.75, slope $p = 0.006$, $r^2 = 0.93$.

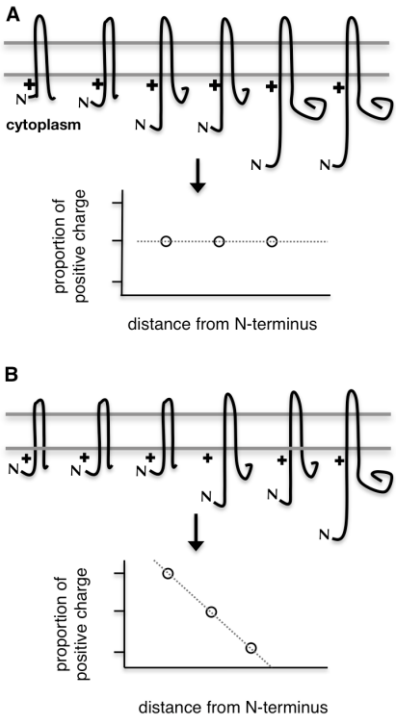

Figure 6. Illustration of how bias in where N-termini cross membranes coupled with positive charge use near the cytosolic leaflet can cause increasing positive charge use nearing the N-terminus. Only a single positive charge is shown on each protein for sake of diagrammatic clarity. **A.** Each of three N-terminal lengths is equally represented, leading to a slope of zero on the resulting regression line. **B.** The locations of where N-termini exit the cytosol are skewed such that the shorter N-terminal length is overrepresented and the longer N-terminal length is underrepresented. This causes an apparent increase in positive charge use at the N-terminus within the "average protein".

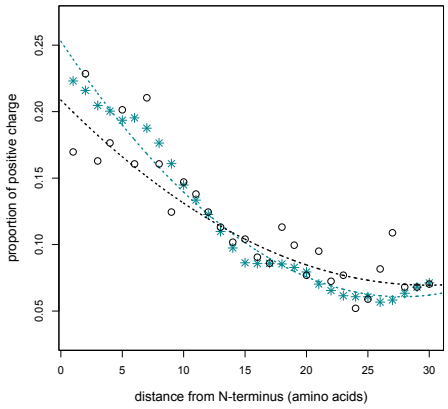

Figure 7. The N-terminal positive charge pattern in *E. coli* can be entirely explained by patterns of positive charge usage near transmembrane regions. Asterisks: N-terminal positive charge reconstructed according to patterns of positive charge usage seen in the vicinity of transmembrane segments which are further downstream (see Results). Regression of $y \sim x^2+x$, x-term coefficient -0.014, p = 5.6e-16. Circles: observed positive charge in cytosolic N-termini (as in Figure 3A, first panel) plotted for comparison. Regression of $y \sim x^2+x$, x-term coefficient -0.0093, p = 5.0e-06, r^2 = 0.82. The observed and reconstructed positive charge usage are not significantly different (paired T-test: p = 0.78; the differences between paired observed and reconstructed charge are normally distributed: Shapiro test, p = 0.083).

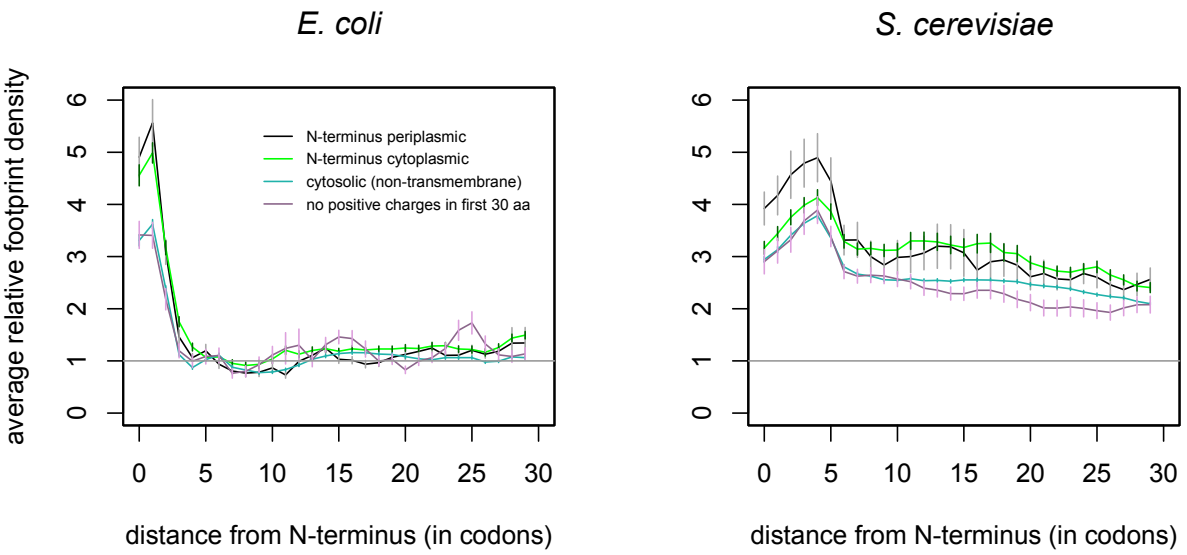

Figure 8. Relative ribosomal occupancy at the beginning of transcripts. Average within-transcript relative ribosomal occupancies were calculated for different subsets of genes as described in Methods, "Ribosomal footprint data". In short, the *y*-axis represents the changes in ribosomal occupancy from one codon position to the next *relative to the occupancy average per site of that transcript*, these relative values then being averaged over aligned transcripts. The gray line at y=1 represents the point at which the ribosomal occupancy in a given position is equal to the average ribosomal occupancy per site of that gene. In all plots, the most increased ribosomal occupancy is seen at the start (approximately the first 4-6 codons or 12-18 nucleotides) of transcripts. ***E. coli.*** Excess occupancy is seen in all categories, but particularly among transmembrane proteins, but strictly only for the first ~4 codons of a transcript. That the relative ribosomal densities return to the ribosomal occupancy average ($y = 1$) after just a few codons, for all protein categories, strongly suggests this initial ribosomal excess is an initiation artefact (see also Discussion). For the rest of the gene ($x > 4$), standardized major axis regression test that y~x slope is not different from 0 from $4 > x < 30$, $p < 2.2\text{e-}16$ with positive (i.e. increasing approaching $x = 30$) slopes given for all plotted categories, contradicting the downward slope that a ramp would predict. ***S. cerevisiae.*** Excess occupancy is, somewhat similarly to *E. coli*, particularly enriched in all categories at the extreme 5' end (up to about $x = 6$), but even after this, occupancy is visibly enriched above the gene average and continues to decrease along the length of the plot.